Why KIRA-EMTA™ Multi-Threaded Supercomputer

The A3Cube™ KIRA-EMTA™ supercomputer realizes an adaptive supercomputing architecture that provides extreme scalability, sustained performance, unique resources convergence, and elastic use of memory devices.

Inside KIRA-EMTA™

KIRA-EMTA™ system leverages the advantages of A3Cube enhanced Dragonfly-Hypercube and Flattened Butterfly combined network topology, InfiniBand interconnect, latest Intel® Xeon® Knights Landing (KNL) processors, integrated convergent parallel storage, elastic caching solutions and latest A3Cube Fortissimo Foundation™ OS suite and programming environment.

- Sustained and scalable application performance
- Tight HPC optimization and integration
- Upgradability by design
- User productivity

Elastic Supercomputing Architecture

Understanding that no single processor engine and static system configuration is ideal for every type of user application, the KIRA-EMTA™ supercomputer platform highlights the flexibility of scalar processing, coprocessing and accelerators, shared memory OS cooperative-kernels, elastic memory, and converged parallel data and IO access, to build a new generation of hybrid systems capable of leveraging the strengths of each technology into one “elastic” HPC environment.
**Intel® Knights Landing Processor**

**Cores**
X86 ISA binary compatible
4 threads / core
Threading: back-to-back fetch and issue per thread
Core resources dynamically repartitioned (shared) between threads at thread selection points

**KNL Instruction Set (ISA)**
First processor that supports **AVX-512**
(512 bit Vector Instructions)
Binary compatible with Intel® Xeon®: Prior Intel® Xeon® processor binaries will run on KNL without recompilation

Both native and offload programming paradigms supported

---

**A3Cube’s Enhanced Versatile Topology network**

- Ultra-low latency
- Extreme IO performance
- Nearest-Neighbors

- **64** nodes max **2** hops (Rack level)
- **1024** nodes max **3** hops (16 Racks)
- Island latency <690ns
- Max latency 128 nodes <800ns
- Max latency 129-1024 nodes 890ns

- **100** Giga bit/s x PE (200 Gbit/s Dual-Rail)
- **2/4** Tera bit/s Island to Island Bandwidth
- **7.2/14.4** Tera bit/s Global Bandwidth
- Up to **400** Millions MPI messages/seconds per PE
- Up to more than **14** Billions MPI Messages/second Routing Capability

---

© A3CUBE Inc.
http://www.a3cube-inc.com
A3Cube Customized ROUTER

Each router provides both “intra-group” (Computing-Island) links that connect it to other computing nodes in its group and “inter-group” links (also known as global links) that connect it to other groups. The router in a group pools its inter-group links, enabling each group to be directly connected to all the other groups.

A3Cube’s (CDMK™)
Collaborative Distributed Multi Kernels (CDMK™)
An unique, non-coherent, elastic remote memory sharing @ cluster level over the fabric.

Fortissimo Foundation™
Combines distributed elements, memory hierarchy optimization, high speed data processing, building an ALL IN ONE high performance, high parallel and fully converged system. It includes:

- FortissimoFS engine
  Unique convergent parallel file system engine with cooperative caching that provides very high-performance parallel access to multi-server-based disk storage, and guarantees efficient use of the available resources to parallel and distributed applications

- Aggressive distributed memory caching accessible by any application (Ultra Cache)

- Configurable system-wide replication to survive hardware failure

- RamStor™
  An advanced, unique feature of FF capable of putting a real parallel, scalable file system completely in-memory

- Cluster-wide distributed scheduler

Enables in-network computing through scalable hierarchical aggregation protocol technology:
  - 7Tb/s of non-blocking bandwidth
  - 200 Millions of messages per port

Executes MPI collectives operations in the network
  - 10X performance improvement for MPI and SHMEM/PGAS applications

Operates as a Co-Processor for A3Cube’s Collaborative Distributed Multi Kernels (CDMK™) and RAM Over Fabric

Built-In A3Cube parallel IO architecture

Along with each CPU is an high performance IO optimized network connection, data and memory bus run in parallel to provide unmatched high performance data access to all the applications running in KIRA-EMTA™

© A3CUBE Inc.
http://www.a3cube-inc.com
Basic System Configuration & Features

64 Processing Elements

- Computing
- Memory
- Storage
- Fabric

12 TByte
>600 Gbyte/s Bandwidth
>7.2 Tbit/s Global Bandwidth

KIRA-EMTA™

Up 64 Processing Elements (PE) (per single rack)
Intel® Xeon Phi x200 (KNL) processor
Up to 4608 Cores per rack
Up to 18,432 Threads
From 170 to 221 Teraflop per rack (double precision)

Starting from 192 GBbyte RAM memory per node
Starting from 12 TByte RAM memory in one rack
A3Cube’s Collaborative Distributed Multi Kernels (CDMK™) (Different Kernels share object using a shared memory mechanism transparent to the application)

Converged storage (starting from 128 TB per racks raw capacity)
Built-in storage in-line data compression
Parallel data access, starting @ >600 GByte/s I/O capability per rack
A3Cube 1.68 TByte of direct DRAM-I/O write caching (with stream based In-line compression)

Up 200 Gbit/s x PE intra-group link bandwidth and up 4 Tbit/s inter-group (global link) bandwidth
7.2 Tbit/s global bandwidth per rack, with end-to-end QoS and congestion control
400 Millions MPI messages /seconds per PE (14 Billions messages routing capabilities /seconds)
Message Passing Interface (MPI) offload

4096 Cores per Rack*
16,384 Threads*
12 TB RAM*
100 Gbit/s per PE*
200 Millions Messages/Second per node*

* Standard Configuration

Built-in fully graphic user interface

© A3CUBE Inc.
http://www.a3cube-inc.com
**KIRA-EMTA™** is designed to run efficiently both memory intensive, IO-intensive and computational intensive applications in the same machine. The fabric topology is realized to extol the application maximum performance dramatically improving the communication efficiency at the entire system level.

**Climate/Weather**  
**Geophysics**  
**CFD**  
**Energy**  
**Financial Services**  
**Molecular Dynamics**  
**Computational Chemistry**  
**Bioinformatics**

- **Memory Intensive**  
- **Computation/IO Intensive**

---

**Fortissimo Foundation™**

- **Operating System:** Linux Based (kernel 4.4.x long term)
- **Exa-Converged Scale-out Distributed File System**
- **Automatic Resource Discovery**
- **Dynamic Workload Distribution**
- **Automatic Load-balancing**
- A3Cube’s Collaborative Distributed Multi Kernels (**CDMK™**)
- A3Cube’s RamStor™
- A3Cube’s Complete Integrated Management Software (system, storage, memory, fabric)

**Complete Computing Tools** (MPI, SHMEM …)

- MVAPICH optimized for KIRA-EMTA fabric (MPI 3.1)
- Support for OpenMPI (MPI3.x)
- Offloader for collectives communication from MPI process onto KIRA-EMTA interconnect hardware
- Message-based and streaming applications accelerator
- OpenSHMEM
- Support for Intra-node shared memory communication
- Intel® KNL software compilers and toolkits

---

© A3CUBE Inc.  
http://www.a3cube-inc.com
KIRA-EMTA™ MVAPICH

KIRA-EMTA™ optimized MVAPICH provides advanced MPI 3 features (such as User Mode Memory Registration - UMR, Dynamic Connected Transport - DCT and Core-Direct-based support for Non-blocking collectives (Hardware supported by KIRA Interconnect)).

KIRA-EMTA™ incorporates hybrid designs that use Shared memory, and IB channels over hybrid Dragonfly-Flattened Butterfly topology that guarantee latency under 1us in the worst case up to 1024 nodes.

KIRA-EMTA™ MVAPICH implementation provides a unified high-performance runtime that supports both MPI and PGAS programming models on KIRA-EMTA™ clusters.

It enables developers to port parts of large MPI applications that are suited for PGAS programming model.

The unified runtime also delivers superior performance compared to using separate MPI and PGAS libraries by optimizing the use of network and memory resources in combination with A3Cube distributed parallel file system engine (FortissimoFS™)

KIRA-EMTA™ MVAPICH supports:
Unified Parallel C (UPC)
UPC++
OpenSHMEM
Coarray Fortran (CAF) as PGAS models

It can be used to run:

- Pure MPI applications,
- MPI+OpenMP application
- Pure UPC, pure UPC++
- Pure OpenSHMEM
- Pure CAF
- Hybrid MPI(+OpenMP) + PGAS applications

KIRA-EMTA™ Intel® KNL Processor explained

What is, and what is not

Knights Landing is NOT Knights Corner.
Knights Corner, the first Intel Xeon Phi product, was a coprocessor. Knights Corner has been extraordinarily successful powering many of the world’s fastest computers. Nevertheless, Knights Corner required a hot processor, shuffling of data over the PCIe bus, and often the use of offload-style programming due to limited memory capacity and strong Amdahl’s Law effects when running less parallel code.

Knights Landing is a full-fledged, highly scalable, Intel processor. This processor can reach unprecedented levels of performance and parallelism, without giving up programmability. You can use the same parallel programming models, the same tools, and the same binaries that run today on other Intel processors. Programming languages that work for processors, just work for Knights Landing too. Programming models, like OpenMP, MPI and TBB, just work for Knights Landing also.

Restrictive models tailored for GPUs, including kernel programming in CUDA and OpenCL, do not apply to processors, it does not need them.

Knights Landing is NOT going to invalidate prior processor coding efforts

Knights Landing is flexible

Unprecedented configurability allows it to be operated in different “memory modes.” MCDRAM can either be treated as a high bandwidth memory-side cache, or it can be identified as high bandwidth memory, or a little of each. Knights Landing also supports different “cluster modes,” allowing it to behave as a cluster with one, two or four NUMA domains per processor, making easy the code performance optimization.